一种结合空间特征的图像注意力标注算法改进研究 *

徐守坤 [1†]，周 佳 [1]，李 宁 [1,2]，石 林 [1]

(1. 常州大学 信息科学与工程学院 数理学院，江苏 常州 213164; 2. 福建省信息处理与智能控制重点实验室(闽江学院)，福州 350108)

摘 要：针对图像标注和 Attention 机制结合过程中特征选择不充分和预测过程中对空间特征权重比例不足问题，提出了一种结合空间特征的注意力图像标注方法。首先通过卷积神经网络得到图像特征，特征区域与文本标注序列匹配；然后通过 Attention 机制给标注词汇加权，结合空间特征提取损失函数得到基于空间特征注意力的图像标注；最后分别在 Flickr30k 和 COCO 两个数据集上进行验证，通过可视化显示该模型如何自动学习显著区域并生成相应的词汇输出序列。实验结果表明，该方法能较好地提取注意力区域并给出标注，与其他模型对比能够得到更好的标注结果。

关键词：视觉注意力；图像标注；空间特征

中图分类号：TP391 doi: 10.3969/j.issn.1001-3695.2017.08.0869

Improved algorithm for image attention annotation combined with spatial features

Xu Shoukun[1†], Zhou Jia[1], Li Ning[1, 2], Shi Lin[1]

(1. *School of Mathematics & Physics*, *School of Information Science & Engineering*, *Changzhou University*, *Changzhou Jiangsu* 213164, *China*; 2. *Fujian Provincial Key Laboratory of Information Processing and Intelligent Control* (*Minjiang College*), *Fuzhou* 350108, *China*)

Abstract: Aiming at the problem of insufficient feature selection and lack of spatial feature weight in the process of image annotation and Attention mechanism, this paper proposed a method of attention image annotation combined with spatial feature. Firstly, it obtained the image feature by convolution neural network, and matched the feature region with the text label sequence. Then, it used the Attention mechanism to weight the annotation vocabulary, and combining the spatial feature to extract the loss function, the image annotation based on the spatial feature attention. Finally, the Flickr30k and COCO validated on the data set to show how the model automatically learns the salient regions and generates the corresponding vocabulary output sequences. The experimental results show that the method can extract the attention area and give the annotation, and compare with other models can get better labeling results.

Key Words: Visual attention; image annotation; spatial feature

0 引言

机器翻译中序列到序列、编码器解码器框架的成功应用[1]为图像标注领域提供更好的实现和使用。Kiros 等人[2]提出一种多模态对数双线性模型前馈神经网络预测下一个标注词汇。Vinyals 等人[3]使用 LSTM 代替 RNN 作为解码器，最后使用 CNN 全连接层输出图像标注。Karpathy 等人[4]将物体检测结果从 R-CNN 和双向 RNN 输出，得到标注排序和联合嵌入空间。近些年注意力机制被引入编码器解码器神经框架得到了更好的图像标注效果，注意力机制由机器翻译发展而来，将人类神经注意力因素考虑到对图像的标注中使得图像中的信息更好的被提取和标注。Xu 等人[5]将注意力机制用于生成相应的图像对齐词汇，提出了基于 LSTM 模型的隐状态结合视觉注意力的模型。该模型也是目前发展比较成熟的基于注意力的图像标注模型。Yang 等人[6]扩展当前注意力编码器解码器框架，加入验证网络，用向量捕捉全局属性加入解码器机制。You[7]和 Wu 等人[8]使用 LSTM 的输入或输出来处理语义几何图像视觉注意力属性的问题，也得到了不错的效果[9,10]。

本文方法主要通过卷积神经网络训练提取且对网络中的空间特征因子权重增加提取，使用 Attention 机制的 LSTM 模型作为编码器解码器，以注意力加权结合空间特征进行图像标注，得到基于空间特征注意力的图像标注结果，最后通过可视化展

基金项目：闽江学院福建省信息处理与智能控制重点实验室开放课题（MJUKF201740）

作者简介：徐守坤（1972-），男（通信作者），吉林蛟河人，教授，博士，主要研究方向为人工智能、普适计算等（zjjuly@163.com）；周佳（1991-），女，硕士研究生，主要研究方向为自然语言处理、图像识别；李宁（1974-），男，副教授，博士，主要研究方向为数据与信息处理；石林（1979-），男，副教授，硕士，主要研究方向为数据处理、图像识别．

示注意力权重与图像标注结果及其分析。

1 相关概念

1.1 Attention 编码器解码器

Attention model（注意力模型）是一种模拟人脑注意力的模型，其核心为 Encoder-Decoder 过程。Encoder-Decoder 模型是一种经典的自然语言处理模型，主要是通过 Encoder 模块对于输入序列进行编码得到编码之后的 code，然后将 code 输入到 Decoder 模块进行解码，最后输出特定的序列。图 1 给出了 Encoder-Decoder 模型的一般框架。

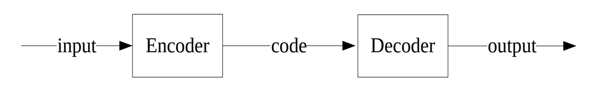


图 1 编码器解码器模型框架

Input 一般是序列 $X=\{x_1,x_2,x_3,\ldots x_n\}$，output 是序列 $Y=\{y_1,y_2,y_3,\ldots y_m\}$。在 Encoder 模块对输入序列进行编码，用 C 表示编码之后的 code，表达公式为：$C=F(x_1,x_2,x_3,\ldots x_n)$。在 Decoder 模块对 C 解码，计算输出 y_i 要用到 C 和之前生成的 $y_1,y_2,y_3,\ldots y_{i-1}$，计算公式为：$y_i=G(C,y_1,y_2,y_3,\ldots y_{i-1})$。由此可以看出在 Decoder 模块中计算输出 y_i 时，用到的语义信息都是一样的，对于较长序列的输入，由于语义编码 code 向量的维度限制，部分有效信息被丢失。

引入的 Attention Model 机制原理在 Decoder 阶段计算出输入序列 $x_1,x_2,x_3,\ldots,x_n$ 对于当前输出的 y_i 的注意力概率分布，对应唯一的语义编码信息，这种编码信息融合了输入对当前输出的注意力概率分布，可以优化当前的输出。加入 Attention Model 的 Encoder-Decoder 模型的框架示意图如图 2 所示。

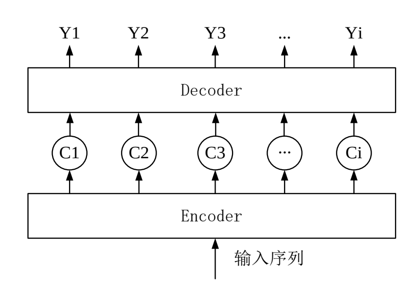


图 2 AM 框架

从图 2 中可以看出，每个输出元素都有对应输入序列概率分布的语义编码 C。因此对于输出 y_i，可以得到这样的计算公式；$y_i=F(C_i,y_1,y_2,y_3,\ldots y_{i-1})$。其中：$C_i$ 是对输入序列 $x_1,x_2,x_3,\ldots,x_n$ 在编码阶段进行非线性函数转换得到；对于输入 $x_i,S(x_i)$ 表示在编码阶段经过函数处理之后的值。编码阶段得到输入序列对应状态值，然后计算出状态值对于输出 y_i 的注意力概率分布。再根据注意力概率分布计算出对应 C_i。计算公式为：$C_i=\sum_{j=1}^{T}\alpha_{ij}S(x_j)$。其中：$\alpha_{ij}$ 是输入 x_j 对输出 y_i 的注意力概率；T 为输入序列的元素的数目。这样设计的原理是计算出 $x_1,x_2,x_3,\ldots x_T$ 对总体的影响力权重，可以突出关键词的作用，减少非关键词对于整体语义的影响。

将 Attention 机制加入 Encoder-Decoder 有两个计算过程分别为计算注意力概率分布下的语义编码及特征向量。具体计算步骤如下：

a)计算注意力分布概率的语义编码，主要思想是计算历史节点和最后输入节点的关系分数，然后计算占总体分数的比重，通过以下公式得到了每一个输入对于最后输入的注意力概率。计算公式如下：

$$a_{ki}=\frac{\exp(e_{ki})}{\sum_{j=1}^{T}exp(e_{kj})} \tag{1}$$

$$e_{ki}=v\tanh(Wh_k+Uh_i+b) \tag{2}$$

其中：a_{ki} 表示节点 i 对于节点 K 的注意力概率权重；v,W,U 为权重矩阵；h_k 为最后输入对应的隐藏层状态；h_i 表示输入序列第 i 个元素对应的隐藏层的状态值。

b)计算注意力分布概率的语义编码和特征向量。其中语义编码 C 主要是通过注意力概率权重与历史输入节点的隐藏层状态乘积的累加得到。最终的语义编码是将含有历史节点的注意力概率分布的语义编码和文章总体向量作为传统 LSTM 模块的输入，最后节点的隐藏层状态值 $H_{k'}$ 就是最终的特征向量。该特征向量，包含了历史输入节点的权重信息，突出了关键词的语义信息[11]表。计算公式如下:

$$C=\sum_{i=1}^{T}a_{ki}h_i \tag{3}$$

$$H_{k'}=H(C,h_k,X') \tag{4}$$

1.2 空间特征

一般地，用卷积神经网络来抽取图像特征，通过多个串行的卷积层(convolution layer)和池化层(pooling layer)间隔排列的方式逐层学习图像数据特征。采用卷积操作方式利用小于图像尺寸卷积核扫描整个图像并计算卷积核与图像局部位置权重之和。每个卷积都对应一个特征映射，随后被输入到池化层进行空域上子抽样(subsample)，使得卷积神经网络具有一定抗畸变能力。网络最顶层将所有得到的特征映射重新拉成一维向量并结合多分类回归分类器反向传播错误信号来调整网络参数。

空间特征是静态图像中物体目标的空间判断能力的重要部分。图像数据的重要特性即数据在空域(两维)和时域(一维)上都存在着明显的统计相关性。在图像标注领域中大多是使用全特征提取，目前全特征提取存在明显缺陷即数据进入网络拉成一维向量形式。这破坏了空域和时域上的相对位置关系，可能会引起相关信息丢失，会产生图像中目标空间方位判断失误，抽取的特征可能引入了其他无关信息。

通过计算两帧之间的逐元素乘积来抽取时域特征，使用多个并行卷积层抽取特征，再计算这些特征的两两逐元素乘。这种神经元间的乘法交互(multiplicative interactions)模型可以显性地学习到时域动态空间特征，同时保留了卷积神经网络在处理空域特征上的优势[12]。

2　模型构建

2.1　基于 Attention 的编码器解码器整体架构

图 3 所示基于注意力机制的循环网络编码解码整体架构。首先分析和表示图像提取视觉特征的多个视觉区域，然后采用视觉特征经 Attenion LSTM 结构即加入了 Attention 机制编码器解码器的 LSTM 网络来预测不同区域的序列，最后得到基于视觉注意力的标注词语生成序列。该模型可以看成是对高维原始输入数据编码之后再解码成低维抽象特征的过程，通过编码器—解码器框架处理各模块之间的关联[16]。

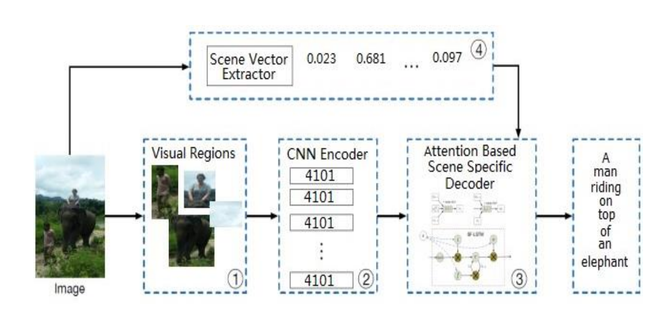


图 3　模型架构

2.2　编码器：卷积特征

模型获取单个原始图像并生成标注编码为 1~K 的编码单词序列。

$$y=\{y_1,\ldots,y_C\},\ y_i\in\mathbb{R}^K \tag{5}$$

其中：K 是词汇表大小，C 是标签长度。使用 CNN 提取作为特征向量的注释向量 a，提取器产生 L 个向量，对应图像的不同空间位置特征用 D 维向量来表示。

$$a=\{a_1,\ldots,a_L\},\ a_i\in\mathbb{R}^D \tag{6}$$

为了获得特征向量和部分图像对应关系，将逐层卷积得到的特征图直接通过全连接层输入到包含 512 个神经单元的下一隐藏层。这使得解码器选择性地聚焦于图像的某些部分，并且加权所有特征向量子集[17,18]。

2.3　解码器：Attention LSTM 网络

将视觉注意力机制引入到网络中，使得每个时刻可以自适应地将注意力集中于当前画面中面积相对较小但具有丰富信息的图像区域，从而加快模型解码速度。使用 LSTM 网络做解码器：

$$\begin{aligned}
i_t&=\sigma(W_iEy_{t-1}+U_ih_{t-1}+Z_iz_t+b_i)\\
f_t&=\sigma(W_fEy_{t-1}+U_fh_{t-1}+Z_fz_t+b_f)\\
c_t&=f_tc_{t-1}+i_t\tanh(W_cEy_{t-1}+U_ch_{t-1}+Z_cz_t+b_c)\\
o_t&=\sigma(W_oEy_{t-1}+U_oh_{t-1}+Z_oz_t+b_o)\\
h_t&=o_t\tanh(c_t)
\end{aligned} \tag{7}$$

其中 i_t,f_t,c_t,o_t,h_t 分别是 LSTM 的输入门、遗忘门、记忆单元、输出门和隐层状态表示；W,U,Z,b 是权重矩阵和偏差；$E\in\Re^{m\times K}$ 是嵌入矩阵；σ 是 sigmoid 函数；上下文向量 $z_t=\sum_{i=1}^{L}\alpha_{ti}a_i$ 是一个动态向量表示在 t 时刻相关部分图像的特征，α_{ti} 表示在时刻 t 中视觉向量 a_i 加权，定义如式(8)所示[20]。LSTM 通过平均注释向量来初始化存储状态和隐藏状态，通过两个分类的 MLPs 得到，如(9)所示。

$$\alpha_{ti}=\frac{\exp(e_{ti})}{\sum_{k=1}^{L}\exp(e_{tk})}\qquad e_{ti}=f_{attn}(a_i,h_{t-1}) \tag{8}$$

$$c_0=f_{init,c}\left(\frac{1}{L}\sum_{i}^{L}a_i\right)\qquad h_0=f_{init,h}\left(\frac{1}{L}\sum_{i}^{L}a_i\right) \tag{9}$$

$f_{attn}(a_i,h_{t-1})$ 是注意力函数，在隐层状态 h_{t-1} 下决定分配给图像特征 a_i 的注意力数量，其中 $\sum_{i=1}^{L}\alpha_{ti}=1$。输出词汇的概率由图像上下文向量 z_t、前一时刻的词汇 y_{t-1} 和隐层状态 h_t 共同决定，如式(10)所示，G 是学习参数。除此对应有损失函数 L，对词汇 $w=\{w_1,\ldots,w_C\}$ 的负采样对数概率，如式(11)所示。

$$p(y_t|a,y_{t-1})\propto\exp(G_o(Ey_{t-1}+G_hh_t+G_zz_t)) \tag{10}$$

$$L_{t,cap}=-\log p(w_t|a,y_{t-1}) \tag{11}$$

兴趣注意力通过模型 $\alpha_t=\{\alpha_{ti}\}_{i=1,\ldots,L}$ 生成。具体来说，正例验证标注的注意力图 $\beta_t=\{\beta_{ti}\}_{i=1,\ldots,L}$ 由正例验证标注给出且 $\sum_{i=1}^{L}\beta_{ti}=1$。一旦 $\sum_{i=1}^{L}\beta_{ti}=\sum_{i=1}^{L}\alpha_{ti}=1$，可以认为是两个注意力概率分布，一般被用在交叉损失熵中验证。对于那些没有与图像区域对齐的单词（如“of”，“is”），设置 $L_{t,attn}$ 为 0：

$$L_{t,attn}=\begin{cases}-\sum_{i=1}^{L}\beta_{ti}\log\alpha_{ti} & \beta\exists w_t\\ 0 & otherwise\end{cases} \tag{12}$$

故而总损失变成两个损失项的加权和：

$$L=\sum_{t=1}^{C}L_{t,cap}+\lambda\sum_{t=1}^{C}L_{t,attn} \tag{13}$$

2.4　空间驱动注意力

空间因素[9]是图像注意力中比较重要的因素，如本文 1.2 节所阐述之原因，故本文将空间特征因素加入注意力模型，从而更好的得到图像标注与生成。CNN 最后一层 ResNet 尺寸为 $2048\times7\times7$，$A=\{a_1,\ldots,a_k\}$，$a_i\in R^{2048}$ 代表全连接层空间卷积特征，k 表示其每个栅格位置，故全局图像特征表示为

$$a^g=\frac{1}{k}\sum_{i=1}^{k}a_i \tag{14}$$

其中：a^g 表示全局图像特征。使用单层感知机和激活函数调整图像特征向量，W_a 和 W_b 是权重参数，得到新的特征向量：

$$\begin{aligned}v_i&=\mathrm{ReLU}(W_aa_i)\\ v^g&=\mathrm{ReLU}(W_ba^g)\end{aligned} \tag{15}$$

图像空间特征最终为 $V=[v_1,\ldots,v_k]$ $v_i\in\Re^d$，每个都用 d 维表示其对应图像部分，故空间注意模型计算 LSTM 的上下文向量 c_t 公式为

$$c_t=f(V,h_t) \tag{16}$$

其中：f 是注意力函数，空间图像特征 $V\in\Re^{d\times k}$ 和 LSTM 隐层状态 $h_t\in\Re^d$，经过单层神经网络由 softmax 函数在图像上得到

包含注意力分布的 k 个区域的空间特征图像：

$$z_t = w_h^T \tanh(W_v V + (W_g h_t) 1^T) \tag{17}$$

$$\alpha_t = softmax(z_t) \tag{18}$$

$1 \in \Re^k$ 是所有元素置为 1 的向量，$W_v, W_g \in \Re^{k \times d} w_h \in \Re^k$ 是学习参数，$\alpha \in \Re^k$ 是在 V 中特征注意力权重。基于注意力分布，上下文向量 c_t 可以如下表示：

$$c_t = \sum_{i=1}^{k} \alpha_{ti} v_{ti} a_{ti} \tag{19}$$

c_t 和 h_t 的组合被用来预测下一个词汇 y_{t+1}。用当前的隐藏状态 h_t 分析注意哪里，结合两种信息源预测下一个词汇。生成上下文向量 c_t 可作为当前隐层状态 h_t 的视觉残差信息，从而减少当前隐层状态预测下一词汇的不确定性。

3　实验及结果分析

3.1　实验设置及评价指标

使用 Flickr30k 和 COCO 两大开源数据集来进行本文实验。Flickr30k 包含 Flickr 收集的 31 783 张图片，图像大多描述了人类日常活动都已被人工标注，每个图像对应五句标注描述。COCO 是目前使用最多的图像标注数据集，包含 82 783、40 504、40 775 幅图像，分别用于训练、验证、测试。因全部图像训练时间过长，所以随机抽取其中一部分融合起来作为实验数据集。将数据集分为三部分：4 000 幅的训练图像、500 幅的验证图像以及 500 幅测试图像，同样每个图像对应有五句人工标注。验证图像主要用于确定模型参数，待参数确定后，验证集里所有图像放入训练集中[21,22]。实验平台为 HP 台式机，硬件配置为 3.2 GHz 的 Intel i5 CPU、4.0 GB 内存，操作系统为 Ubuntu 14.0，软件环境为 MATLAB 2014a 及 python 2.7。图像标注生成常用的评价指标为 BLEU[25]值，本实验依此进行评价，除 BLEU 之外，另一种常见评价指标 METEOR[23]和 CIDEr[24]。评估 Flickr30k 和 COCO，与现有的 MSM[26]，Hard-Attention[5]以及 DeepVS [4]进行部分比较。

3.2　实验结果分析

所有实验情况参数设置等细节严格遵守 Xu 等人[5]的模型。本文调整图像的大小短边为 256 像素，中心区域裁剪成 224×224 像素。预训练 ImageNet 之后提取 VGG19 网络中 conv5_4 特征，顶层卷积层尺寸为 14×14。为了可视化注意力模型权重，上采样权重因子为 2^4=16，使用高斯滤波器模拟感受野大小。设置 CNN 卷积迭代次数为 15 000 次，训练文本向量矩阵迭代次数为 15 000 次；为了避免过度拟合，设置 CNN 的权重下降速率为 10^{-3}。LSTM 语言模型的学习率为 4×10^{-4}；设置更新权重参数为 α =0.8、β =0.999。进行随机梯度下降非正则化训练，为数据源 Flickr30k 设置 1 300 个 LSTM 单元，COCO 数据集为 1 800 个。

在 Caffe 框架下使用 Zhang 等人[26]提供的开源代码训练程序得到的效果如图 4 所示。基于注意力模型的图像语义生成标注正样例，在图中重点注意力在于三个人的特征注意，颜色越深的地方表示注意力权重越深，故而“man”和“boy”等词汇权重较其他来说略高一些。两两人物关系因加入空间特征，故而在同时注意区域推测可能为夫妻关系；男性注意力在男孩身上，故而推测为父亲，推测的标注词汇来源于训练集中的现有语句词汇逻辑预测成果。

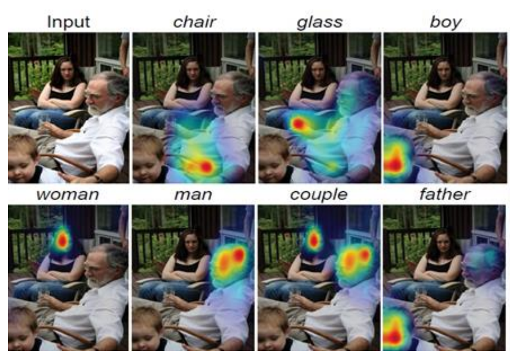


图 4　实验正例

图 5 为实验结果对比。加入了空间因素，可以看出对于注意力权重和空间判断有了矫正。图 5 展示在同一个交通标志在三种现有模型下的注意力对比。对于 STOP 标志的注意力权重比 Hard Attention 模型范围测距小，因加入的空间特征对无关特征去除及加深相关特征权重的缘故，所以比起 DeepVS 是全特征识别更为典型的代表，对于颜色空间过度识别。



图 5　实验比较

进一步将标注生成对应空间注意力关系可视化对应如图 6 所示。非注意力词汇例如“of”对其进行注意力关注提升，因为冠词之后很有可能给出重点名词，如同“riding”“elephant”之类词汇分配注意概率占重比非注意力词汇大。上下文场景环境不同时，同词汇分配的视觉注意概率度也是不同的。例如词汇“a”通常在文章的开始具有较高标注概率，因无背景上下文需要 LSTM 保存信息再判断。

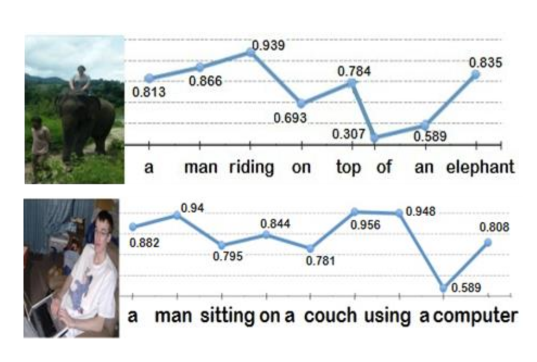


图 6　词汇生成过程

表 1　注意力模型比较/%

模型	Flickr30k						MS-COCO						平均正确率	
	B-1	B-2	B-3	B-4	M	C	B-1	B-2	B-3	B-4	M	C	Flickr30k	MS-COCO
DeepVS	0.572	0.367	0.241	0.158	0.154	0.248	0.628	0.451	0.322	0.231	0.198	0.678	0.627	0.911
Hard-Attention	0.668	0.438	0.286	0.189	0.187	0.185	0.719	0.547	0.358	0.251	0.231	-	0.643	0.906
MSM	-	-	-	-	-	-	0.725	0.561	0.423	0.326	0.250	0.987	0.551	0.923
Ours	**0.671**	**0.442**	**0.257**	**0.190**	**0.194**	**0.255**	**0.731**	**0.570**	**0.334**	**0.331**	**0.257**	**0.990**	**0.657**	**0.925**

如表 1 所示的结果，对 Flickr30k 和 COCO 的数据集，使用 M 表示 METEOR 指标，用 C 表示 CIDEr 指标，使用 Ours 代表本文实验。与没有加入空间特征注意力模型的算法进行比较，本文融入空间注意力模型局部性能上稍优于其他注意模型。在 Flickr30k 数据集上，CIDEr 得分值从 0.248 提高至 0.255；在 COCO 数据集上为从 0.987 提升至 0.990。在 COCO 数据集中，本文方法 BLEU-4 得分从 0.326 提高到 0.331，METEOR 从 0.250 到 0.257。标注模型所 BLEU 指标相比基线有所提升。模型在 Flickr30k BLEU-1 评分提升(0.671-0.668)/0.668 ≈ 0.4%；COCO BLEU-2 评分提升(0.570-0.547)/0.547 ≈ 4.2%。从表 1 可以看出，在准确率方面本文方法有较好的标注效果，在 Flickr30k 数据集下，经过训练和随机局部结果抽样比对提升了 19.2%；MS-COCO 数据集下提升有 2.1%。

对上述模型计算复杂度比较，本文将各模型在两大数据集中随机抽样测试 1 000 张，长度为 20 个字符以内的单张图像对平均标注时间进行比较，结果如表 2 所示。在表 2 中，本文算法运行所需时间相比其他模型运行时间增加的幅度在 0.039~0.320 s，在 Flickr30k 数据上相对增量为 0.11%，在 MS-COCO 上为 0.02%，整体平均增量为 0.07%，尚未达到 1%，可见本文算法复杂度的增加在可承受的范围内。综合评分指标，模型的整体基于空间注意力融合标注局部性能上较优，它具备一定实用价值。

表 2　数据集上平均运行时间/s/张

模型	DeepVS	Hard-Attention	MSM	**Ours**
Flickr30k	2.039	2.217	2.864	**2.359**
MS-COCO	4.483	4.334	5.847	**4.522**

4　结束语

本文在以前工作的基础上提出了一种有效针对图像的视觉融入空间特征注意力模型，能够很好地描述图像中吸引注意力区域的情况。首先通过卷积神经网络得到图像特征，特征图像区域标注匹配，使用 Attention 机制的 LSTM 模型作为编码器解码器，以注意力加权结合空间特征进行图像标注，最终得到基于空间特征注意力的图像标注生成结果。实验结果表明，与相关方法相比，本文所提出的算法在标注性能上取得了一定的效果，但是从整体评估和独创性方面来看还需要很多的改进工作。

参考文献：

[1] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [J]. Computer Science, 2014, 40 (12): 4751-4759.

[2] Kiros R, Salakhutdinov R, Zemel R. Multimodal neural language models [C]// Proc of International Conference on Learning Representations. 2014: II-595.

[3] Vinyals O, Toshev A, Bengio S, et al. Show and tell: a neural image caption generator [J]. Computer Science, 2015, 36 (7): 3156-3164.

[4] Karpathy A, Li F F. Deep visual-semantic alignments for generating image descriptions [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2014, 39 (4): 664-676.

[5] Xu K, Ba J, Kiros R, et al. Show, attend and tell: neural image caption generation with visual attention [J]. Computer Science, 2015, 58 (12): 2048-2057.

[6] Yang Z, Yuan Y, Wu Y, et al. Review networks for caption generation [C]// Advances in Neural Information Processing Systems. 2016: 2361-2369.

[7] You Q, Jin H, Wang Z, et al. Image captioning with semantic attention [J]. Computer Science, 2016, 42 (13): 4651-4659.

[8] Wu Q, Shen C, Liu L, et al. What value do explicit high level concepts have in vision to language problems? [J]. Computer Science, 2016, 12 (01): 1640-1649.

[9] Lu J, Xiong C, Parikh D, et al. Knowing when to look: adaptive attention via a visual sentinel for image captioning [J]. International Journal of Computer Vision, 2016, 115 (3): 211-252.

[10] Zhou L, Xu C, Koch P, et al. Watch what you just said: image captioning with text-conditional attention [J]. IEEE Trans on Image Processing, 2016, 25 (8): 3919-3930.

[11] 张冲. 基于 Attention-Based LSTM 模型的文本分类技术的研究 [D]. 南京：南京大学, 2016.

[12] 杨格兰，邓晓军，刘琮. 基于深度时空域卷积神经网络的表情识别模型 [J]. 中南大学学报：自然科学版, 2016, 47 (7): 2311-2319.

[13] 李静. 基于多特征的图像标注研究 [D]. 武汉：武汉理工大学, 2013.

[14] 滕飞，郑超美，李文. 基于长短期记忆多维主题情感倾向性分析模型 [J]. 计算机应用, 2016, 36 (8): 2252-2256.

[15] 刘杰. LSTM 神经网络在 Android 平台上的实现 [D]. 天津：南开大学,

2015.

[16] Fu Kun, Jin Junqi, Cui Renpeng, et al. Aligning where to see and what to tell: image captioning with region-based Attention and scene-specific contexts [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2015, 39 (12): 2321-2334.

[17] Cho K, Merrienboer B V, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [J]. Computer Science, 2014, 45 (18): 4913-4921.

[18] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks [J]. Advances in Neural Information Processing Systems, 2014, 4 (3): 3104-3112.

[19] Mao J, Xu W, Yang Y, et al. Deep Captioning with multimodal recurrent neural networks (m-RNN) [C]// Proc of International Conference on Learning Representations. 2015: II-301.

[20] Liu C, Mao J, Sha F, et al. Attention correctness in neural image captioning [C]// Proc of AAAI -the Association for the Advance of Artificial Intelligence. 2017: 4176-4182.

[21] 柯逍，李绍滋，曹冬林. 基于相关视觉关键词的图像自动标注方法研究 [J]. 计算机研究与发展, 2012, 49 (4): 846-855.

[22] Denkowski M, Lavie A. Meteor universal: language specific translation evaluation for any target language [C]// Proc of Workshop on Statistical Machine Translation. 2014: 376-380.

[23] Vedantam R, Zitnick C L, Parikh D. CIDEr: consensus-based image description evaluation [J]. Computer Science, 2014, 9 (4): 4566-4575.

[24] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation [C]// Proc of Meeting on Association for Computational Linguistics. 2002: 311-318.

[25] Yao T, Pan Y, Li Y, et al. Boosting image captioning with attributes [J]. ACM Trans on Graphics, 2016, 27 (3): 1423-1436.

[26] Zhang J, Lin Z, Brandt J, et al. Top-down neural attention by excitation backprop [C]// Proc of European Conference on Computer Vision. [S. l.] : Springer International Publishing, 2016: 543-559.